素材再利用のための効果的なメタデータの付加方法に関する研究

千葉華子†　藤澤公也†
東京工科大学大学院バイオ情報メディア研究科†

はじめに

近年のパーソナルコンピュータとインターネットの普及により、大学などにおいて情報の電子媒体化が進んでいる。これに伴い、PC を利用して授業を進める設備も整い、プレゼンテーションソフトを使用して授業を進めるなど電子媒体で授業資料を作成することが増えている。

現在、大学など高等教育機関では広く何度も扱われるような授業資料の再利用が求められている。ただし、単に共有・再利用をしても部分的な利用が難しいことや授業間で共通点の見過ごしなどの問題が多い。この解決のため先行研究[1]として教材開発支援システムの一部である素材管理・分類支援システムの設計・実装を行った。

本研究では、先行研究では解決されなかった内容的な関連性を表すために効果的なメタデータを半自動的に付加する手法の提案を目的とする。

素材管理・分類支援システムの問題点

先行研究では授業資料を PowerPoint(PPT)に限定し、半自動的に親子関係を付加した。しかし、素材間に一定の関係しか付加できず、ある程度の指標にはなるものの内容的な関連を検索する場合に必ずしも正しい結果が出るわけではないという問題があった。また、キーワードも素材の持つテキスト情報を利用したが、PPT のスライド単位では、単語の数が少なく、検索しても欲しい素材が見つからないことが考えられる。人手でキーワードを付加することで、より良い検索結果を得ることができるが、資料のページ数や素材の数などから敬遠されてしまうという問題がある。

新しい関連性の付加

先行研究で付加したメタデータだけでは検索で再利用する資料探すためには情報が足りないため、本研究では素材の再利用時に行われる検索のために本来素材が持っていないキーワードの追加を半自動的に行う。PowerPoint のスライドでは文章量が少なく、その文章からキーワードを付加しただけでは、関連のある資料など検索できない。そこで新しい関連性として、サブキーワードの付加を行い、検索に対応できるようにする。このサブキーワードはキーワードの関連付けを基に素材に付加される。

キーワード同士の関連付けというのは一般的には類語辞書を利用した方法がある。しかし、ここで対象としている授業資料では “電子メール” と “Thunderbird” のように一般の辞書では関係付けすることができない単語が多く出現する。そのため、本研究では類語辞書を補完するような手法提案する。ただし、本システムでは類語辞書は使用しない。

キーワードの関連付けは、一つのスライド(図1 のスライド 1)内に 2 つ以上のキーワード(図 1 の “電子メール” と “Thunderbird” 、 “メールアドレス”)が存在する場合に行う。キーワード同士の組み合わせに対して関係値を付加しデータベースに保存する。これにより、辞書を使わずとも、内容的な関連性について求めることができる。本研究ではこのキーワード付加は形態素解析を利用して行った。今回キーワードとして付加したものは、文中の単語で形態素解析の結果、名詞の一部と未知語に分類されたものを利用した。

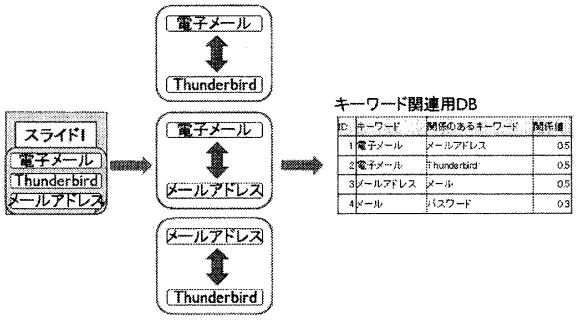

図 1　キーワード同士の関連付け

関係値は同じ組み合わせのキーワード同士が出現する頻度を利用してその値を決定する。

The method of generating metadata for reusing material
†Hanako CHIBA, Kimiya FUJISAWA・Graduate School of Bionics Computer and Media Sciences,Tokyo University of Technology

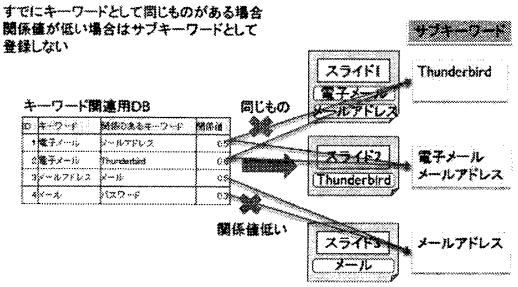

図 2　サブキーワードの付加

そしてキーワードの関連性を用いてサブキーワードの付加を行う。図 2 のようにスライド(図 2 のスライド 1,2,3)のもつキーワードと関係があるキーワードをデータベースから探し、サブキーワードとして素材に付加する。なおサブキーワードとして付加されるものは一定以上の関係を持つキーワード同士とし、これにより不要なサブキーワードの増加を防ぐことができる。

検証

本研究では、キーワード同士の関連付けとサブキーワードの付加をシステムによって行い、実験・検証を行っている。実験には、実際に使用されている授業資料を使って行う。

今回の実験では本学で行われている授業の 4 回分の資料からスライドを選択し、キーワードの関連付けとサブキーワードの付加をシステムによって行った。サブキーワードの付加を行う場合、キーワード同士の出現回数が 4 回以上のものと 2 回以上のものをサブキーワードとして場合の 2 通り行った。

付加されたキーワードをそのまま利用して実験を行った場合、4 回分の資料の持つキーワードの総数が 259 個となった。出現回数が 2 回以上場合と 4 回以上場合、どちらも付加したサブキーワード数が多く、2 桁から 3 桁付加された(表 1)。内容的にも付加されたサブキーワードの内容を調べたところ、無駄なサブキーワードが多かった。特に、出現回数 2 回以上の場合(表 1 の整理前)には 100 前後のサブキーワードが付加され有効なものは少ないという結果になった。

表 1　同じ授業資料 4 回分実験結果

素材ID	キーワード数		サブキー4回以上					サブキー2回以上				
			有効数		有効数		有効数	有効数		有効数		有効数
	整理前	整理後	整理前		整理後		の差	整理前		整理後		の差
33	11	5	42	10	5	4	-6	149	17	15	8	-9
90	15	5	40	10	6	5	-5	120	14	17	9	-5
94	13	6	41	10	7	5	-5	131	13	20	10	-3
126	10	8	38	10	4	3	-7	118	12	16	8	-4
134	11	7	33	4	2	2	-2	124	10	13	5	-5
147	7	3	10	2	1	1	-1	40	4	5	2	-2
149	4	3	36	11	5	5	-6	102	13	16	7	-6
160	5	4	39	11	3	2	-9	127	12	18	9	-3
178	5	3	19	9	5	5	-4	83	13	12	8	-5
186	4	2	18	8	4	4	-4	64	8	11	8	0

この実験の結果から、意味のないキーワードが多数付加されている状況では、サブキーワードのほうも意味のないものが付加されてしまう。そこで上の実験で利用した資料に付加されたキーワードを人の手で、その授業に必要と考えられるキーワードだけになるように整理を行った。このキーワードの整理を行ったものではキーワードの数が 95 個と整理していないものと比べて半分以下になった。この状態で同じ実験を行った。出現回数 2 回以上で 15 前後、4 回以上で 5 個前後のサブキーワードが付加された(表 1)。調整前に比べてサブキーワードの数は減り、4 回以上の場合(表 1 の整理後)は有効とされるサブキーワードの比率が高くなっており良い結果と言える。しかし、内容的な判断を行うと 4 回以上の場合はほとんどの素材に同じサブキーワードが付加されているという状態であり、絞り込みや検索に対しての有効性は期待できないと考えられる。2 回以上の場合は素材ごとに違異なるサブキーワードが付いているので内容的には一番検索に効果的な結果になったと言える。

おわりに

本研究では、授業資料の再利用時に行われる検索で関連性のある素材を検索するために必要なメタデータの付加とその付加方法についての研究を行う。現在、教育用システムにおける必要なメタデータを検討し、それを付加するシステムの実装を行っている。新しく付加する関連性として、キーワード同士の関連付けとスサブキーワード付加の二つを自動的に付加する。そして実験を行い、付加したサブキーワードの有効数や付加数、検索する場合の利点などを考えてどのような付加を行えば有効なサブキーワードの付加を行うことができるか検証し、有効なサブキーワードを付加できる条件を絞り込んだ。

キーワードの付加を行う場合に授業に関連のある意味を持つキーワードの判断を行う方法を提案できれば本研究のサブキーワード付加もよりよい方向に利用できると考えている。

参考文

[1] 柴田ちひろ・千葉華子・藤澤公也, “授業資料作成支援システムの構築：素材 DB 蓄積及び授業内容構造化”, 情報処理学会第 69 回全国大会, pp. 4-701-4-702, 2007